Do Membership Inference Attacks Work on Large Language Models?

Year: 2024 | *Citations: 148* | *Authors: Michael Duan, Anshuman Suri, Niloofar Mireshghallah*

Abstract

Membership inference attacks (MIAs) attempt to predict whether a particular datapoint is a member of a target model's training data. Despite extensive research on traditional machine learning models, there has been limited work studying MIA on the pre-training data of large language models (LLMs). We perform a large-scale evaluation of MIAs over a suite of language models (LMs) trained on the Pile, ranging from 160M to 12B parameters. We find that MIAs barely outperform random guessing for most settings across varying LLM sizes and domains. Our further analyses reveal that this poor performance can be attributed to (1) the combination of a large dataset and few training iterations, and (2) an inherently fuzzy boundary between members and non-members. We identify specific settings where LLMs have been shown to be vulnerable to membership inference and show that the apparent success in such settings can be attributed to a distribution shift, such as when members and non-members are drawn from the seemingly identical domain but with different temporal ranges. We release our code and data as a unified benchmark package that includes all existing MIAs, supporting future work.